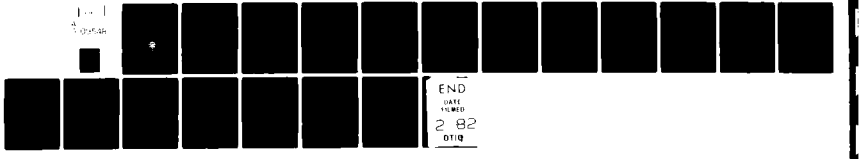
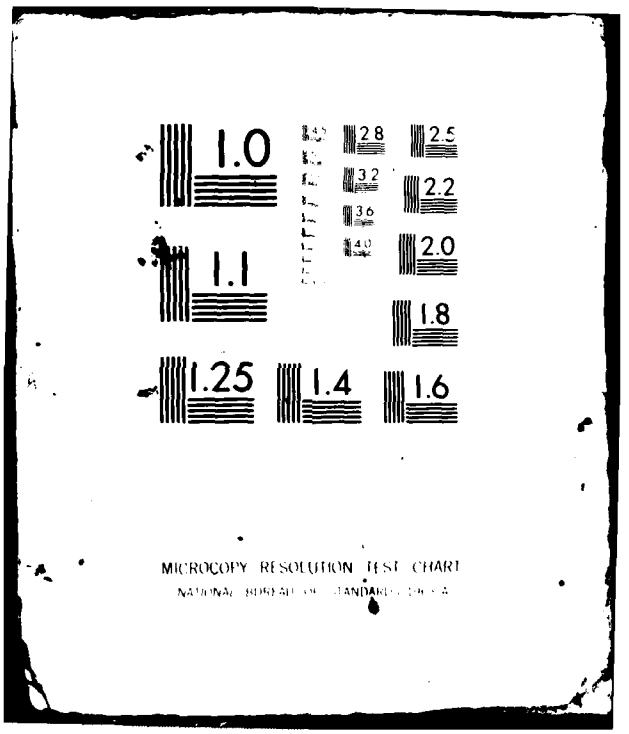
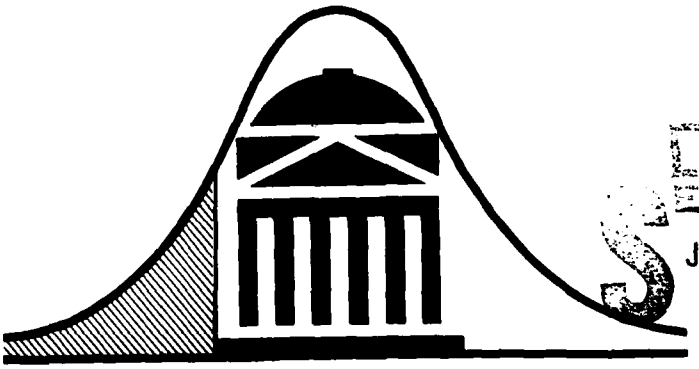
AD-A109 548 SOUTHERN METHODIST UNIV DALLAS TEX DEPT OF STATISTICS F/G 12/1
TESTING THE CORRELATION COEFFICIENT WITH INCOMPLETE OBSERVATION--ETC(U)
NOV 81 J BECKETT, W R SCHUCANY, N J BOSMIA N00014-75-C-0439
UNCLASSIFIED TR-153 NL



END
DATE FILMED
2 82
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART

SOUTHERN METHODIST UNIVERSITY



DEPARTMENT OF STATISTICS

DALLAS, TEXAS 75275

TESTING THE CORRELATION COEFFICIENT WITH INCOMPLETE OBSERVATIONS

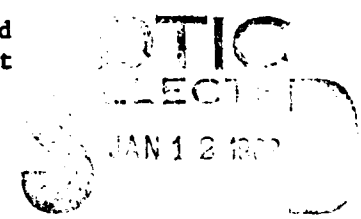
by

James Beckett, William R. Schucany, N. J. Bosmia

Technical Report No. 153
Department of Statistics ONR Contract

November, 1981

Research sponsored by the Office of Naval Research
Contract N00014-75-C-0439



A

DEPARTMENT OF STATISTICS
Southern Methodist University
Dallas, Texas 75275

TESTING THE CORRELATION COEFFICIENT WITH INCOMPLETE OBSERVATIONS

by

James Beckett
Bowling Green State University

William R. Schucany and N. J. Bosmia
Southern Methodist University

ABSTRACT

Correlation is often investigated (and tested for significance) in situations where some of the observations on one of the variables are missing. Throwing away these unpaired observations may seem to be a waste of information; a test based on all the data at hand would seemingly be better than a test based on only some of the data available. An exact test using all the data, which is similar in form and distribution to the usual t test based on the sample correlation coefficient, is derived and examined. However, this exact test proves to be a relatively inefficient way to incorporate the extra information. This counterintuitive result provides an interesting lesson concerning the relationship between power and degrees of freedom.

I. INTRODUCTION

Estimating or testing the correlation between two variables is a common problem with applications reaching into virtually every

subject area which makes use of the science of statistics. In many of these diverse applications, data on one or more of the variables of interest may be lost, missing, or unobtainable for some of the subjects. In a medical setting where physiological measurements are to be obtained, subjects dying, instruments malfunctioning, and other random miscellaneous situations (the occurrence of which should in no way be related to any of the variables or treatments) do lead to a missing data problem.

Practitioners of statistics since Wilks (1932) have realized that there is additional information in the unpaired observations. The question is how to properly and efficiently use this extra information such that the resulting test would be more powerful than the standard t test based exclusively on the paired observations. Herein, a t test (with greater degrees of freedom) is derived, but in spite of the similarity in form, the t test with more degrees of freedom is in fact inferior with respect to power for some alternatives.

We have found this result to be pedagogically valuable on several counts. First, the beginning student can be led down the primrose path for a greater impact of the point that one's intuition about a reasonable way to do things may not be infallible. We consider it an important general lesson that one must take care not to misuse information thinking that this is in some way preferable to not using it at all. Finally, this new statistic is an exception to the rule that more degrees of freedom yields more power when comparing two exact tests of similar form.

This new test is derived in much the same way as the usual t

test for correlation based only on the paired observations. Consider (x_i,y_i), i=1,2,..., m+n to be a random sample of size m+n from a bivariate normal distribution with parameters μ_x, σ_x^2, μ_y, σ_y^2, and ρ. Suppose m of the x values are missing (or that we have m "extra" y values) and that we would like to test the null hypothesis $H_o:\rho=0$ versus $H_1: \rho > 0$. Rearranging the indices for convenience we obtain:

$$x_1,x_2,x_3,\ldots,x_n$$

$$y_1,y_2,y_3,\ldots,y_n,y_{n+1},\ldots,y_{n+m} .$$

The existence of an exact test based on the sample product moment correlation coefficient is well known, using

$$r = \frac{\sum_{i=1}^{n} (x_i-\bar{x}_n)(y_i-\bar{y}_n)}{\sqrt{\sum_{i=1}^{n} (x_i-\bar{x}_n)^2 \sum_{i=1}^{n} (y_i-\bar{y}_n)^2}}$$

where $\bar{x}_n = \sum_{i=1}^{n} x_i/n$ and $\bar{y}_n = \sum_{i=1}^{n} y_i/n$ if there are an equal number of x's and y's, that is, bivariate observations.

In the case at hand (n+m) paired observations are not available. Hence one way to test H_o is to discard the additional unpaired observations. But to do so would be discarding some information. Since the unpaired y's do give information about some of the parameters, it would seem reasonable that we should use these y's in some fashion.

Three tests are investigated. The first one is an exact test. The second one is based on the maximum likelihood estimate of ρ and the third one is based on the generalized likelihood ratio. It will be shown that the generalized likelihood ratio test does not depend

on the additional information.

II. AN EXACT TEST

Let us now examine a test statistic which uses all of the data and is similar in form and distribution to that of the familiar test procedure for complete paired samples. The temptation is to accomodate the extra observations of y in a straightforward manner by defining

$$r^* = [\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_{n+m})]/[\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 \sum_{i=1}^{n+m}(y_i-\bar{y}_{n+m})^2]^{1/2},$$

where $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y}_{n+m} = \frac{1}{n+m}\sum_{i=1}^{n+m} y_i$.

We may consider r* as being derived from the following naive estimator of ρ

$$\tilde{\rho} = \frac{\tilde{\sigma}_{xy}}{\tilde{\sigma}_x\tilde{\sigma}_y} = [\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)]/[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 \frac{1}{n+m}\sum_{i=1}^{n+m}(y_i-\bar{y}_{n+m})^2]^{1/2}$$

$$= [n+m)/n]^{1/2} r^* .$$

We note that these individual estimators do not all coincide with the maximum likelihood estimators which were given explicitly in this case by Anderson (1957); in fact the maximum likelihood estimator of ρ is seen to be something quite different from $\tilde{\rho}$. However, the following theorem concerning the exact null distribution of a test statistic based on r* is quite similar to the result for r, the MLE in the complete sample case.

Theorem

Under the hypothesis H_0: $\rho = 0$, $t^* = r^*(n+m-2)^{1/2}/(1-r^{*2})^{1/2}$ is distributed as Student t with n+m-2 degrees of freedom.

Proof: Define two (n+m) x 1 vectors $\underline{Z}_1$ and $\underline{Z}_2$ with elements

$$Z_{1i} = \begin{cases} x_i - \bar{x}_n & \text{if } i = 1, 2, \ldots, n \\ 0 & \text{if } i = n+1, \ldots, n+m \end{cases}$$

and

$$Z_{2i} = y_i - \bar{y}_{n+m} \quad \text{for } i = 1, 2, \ldots, n+m.$$

Hence $\underline{Z}_2 \sim \text{MVN}(\underline{0}, \sigma_y^2(I - \frac{1}{n+m} J))$, where I is an (n+m) x (n+m) identity matrix and J is an (n+m) x (n+m) matrix of ones. The conditional distribution of $b = (\underline{Z}_1'\underline{Z}_1)^{-1}\underline{Z}_1'\underline{Z}_2$ is normal with mean 0 and variance $(\underline{Z}_1'\underline{Z}_1)^{-1}\sigma_y^2$, given the x's.

Further let

$$V = (\underline{Z}_2 - b\underline{Z}_1)'(\underline{Z}_2 - b\underline{Z}_1) = \underline{Z}_2'\left[I - \frac{\underline{Z}_1\underline{Z}_1'}{\underline{Z}_1'\underline{Z}_1}\right]\underline{Z}_2 .$$

Finally let W be the product of the matrix of the quadratic form, V, and the covariance matrix of $\underline{Z}_2$. Thus W has the form

$$W = \left(I - \frac{\underline{Z}_1\underline{Z}_1'}{\underline{Z}_1'\underline{Z}_1}\right)\left(I - \frac{1}{n+m} J\right) = I - \frac{1}{n+m} J - \frac{\underline{Z}_1\underline{Z}_1'}{\underline{Z}_1'\underline{Z}_1} .$$

The quadratic form V has a chi-square distribution if and only if W is idempotent. The idempotence of W can be easily shown and by inspection the rank (trace) of W is n + m - 2. Thus $V/\sigma_y^2 \sim \chi^2_{n+m-2}$ (conditioned on $\underline{Z}_1$). Now b and V are independent if

$$\frac{\underline{Z}_1'}{\underline{Z}_1'\underline{Z}_1} W = \emptyset .$$

Noting that $\underline{Z}_1$ and J are orthogonal, the independence follows immediately. Therefore $t^* = [b(\underline{Z}_1'\underline{Z}_1)^{1/2}/\sigma_y]/[(V/\sigma_y^2)/(n+m-2)]^{1/2} \sim t_{n+m-2}$ conditional on $\underline{Z}_1$. Rewriting in terms of $\underline{Z}_1$ and $\underline{Z}_2$ yields

$$t^* = r^*(n+m-2)^{1/2}/(1-r^{*2})^{1/2} \quad .$$

Since the conditional null distribution of the quantity, t*, in no way depends on $\underline{Z}_1$(X's), it is the unconditional distribution as well. Q.E.D.

This test is appealing in its simplicity and hence our next concern is its efficiency. There is even some cause for optimism because of the increased degrees of freedom. The theorem and proof above give the null distribution of t*; however the derivation of the non-null distribution of t* is not a simple task. For any simple alternative hypothesis that $\rho = \rho_0 \neq 0$, the distribution of t* is not a non-central t except in the special case where $m = \infty$; and hence the exact power can be calculated in this case. The analytical calculation of the power function in cases other than $m=\infty$ is difficult. However, a small scale sampling experiment is sufficient to demonstrate the bad news...that we would be better off to throw the extra y's away rather than to use them as in t*.

III. Power Comparison

The power functions of these two competing procedures (t^*_{n+m-2} and t_{n-2}) are compared using a small Monte Carlo study. David (1954) has tabulated the distribution of r for different values of n and ρ. Using her tables the true power of the test based on r is given in Table 1 both for comparison with the empirical power of t* and as a validation of the simulation by their close agreement with the empirical power of t.

For the comparison of t_{n-2} with t^*_{n+m-2}, 2000 samples were generated with n x's and n+m y's. In each case, n = 10 while m

TABLE 1

Power of tests of H_o: $\rho = 0$ vs. H_1: $\rho > 0$ ($\alpha = .05$)

Monte Carlo Estimates Based on 2000 Samples

ρ	True Power	Empirical Power			
	t(10,10)	t(10,10)	t*(10,20)	t*(10,50)	t*(10,∞)
0	.050	.054	.047	.050	.052
.10	.085	.096	.082	.086	.082
.20	.138	.146	.136	.142	.135
.30	.215	.228	.211	.213	.210
.45	.390	.393	.374	.367	.357
.60	.623	.620	.587	.560	.536
.75	.867	.860	.792	.764	.739
.90	.993	.994	.946	.921	.897

Maximum Standard Error of Table Entries is .011.

varied over the values 0, 10, 40 and ∞ (limiting case of known μ_y and σ_y). Across columns the samples consist of the same random samples of 10 paired observations plus possibly some additional unpaired y's to stabilize the comparison of procedures. The computations were done on the Univac 1100 computer at BGSU. The random normal deviates were generated using the IMSL library subroutine GGNOR with one of their recommended seeds.

There is no evidence that t* is more powerful than t for any value of m for any alternative value of ρ. On the other hand, t is in fact significantly more powerful than t* for the larger ($\rho \geq .60$) alternatives. Thus t* cannot be recommended over the usual t-test which is based upon only the paired observations.

As $m \to \infty$, we get the smallest power for values of $\rho \geq .45$, which may be surprising considering that as $m \to \infty$ we say we "know" μ_y and σ_y. More and more power is lost (relative to the standard paired procedure) as the true value of ρ departs from the null. The misuse of the extra y's in r* is more and more evident as the underlying true correlation increases and as the amount of unpaired data begins to dominate the paired data. It should be noted that in this limiting case of "known" μ_y and σ_y, t* becomes

$$t^*(n,\infty) = \left[\sum_{i=1}^{n} (x_i-\bar{x})(y_i-\mu_y)/\sigma_y\right] / \left[\sum_{i=1}^{n} (x_i-\bar{x})^2\right]^{1/2}$$

which is distributed normally with mean zero and variance one under the null hypothesis.

The poor performance of even the limiting case of r* is in line with Wilks' (1932) estimation results of five decades ago. Thus, in spite of the fact that r* provides an exact test for $\rho = 0$, it repre-

sents a less efficient use of the extra y values than discarding them.

IV. A TEST BASED ON THE MLE

The simplification available in the case of μ_y and σ_y known lends itself to further examination. We focus attention on this case because the power function has been calculated for the exact test, t*.

The ML estimate, $\hat{\rho}$, is given by Anderson (1957) for unknown μ_x, μ_y, σ_x, σ_y and ρ. From this, the ML estimate for known μ_y and σ_y is easily deduced. Using the following notation

$$S_{xy}^2 = \sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n) ,$$

$$S_x^2 = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$$

and

$$S_y^2 = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 .$$

Solution of the likelihood equations yields

$$\hat{\rho} = \sigma_y \frac{S_{xy}}{S_y^2} \left\{ \frac{S_x^2}{n} + \frac{S_{xy}^2}{S_y^4} \left(\sigma_y^2 - \frac{S_y^2}{n} \right) \right\}^{-1/2} .$$

It can be shown that

$$\frac{\hat{\rho}^2}{1-\hat{\rho}^2} = n\sigma_y^2 \frac{S_{xy}^2}{S_y^2 (S_x^2 S_y^2 - S_{xy}^2)} .$$

Letting $r = \frac{S_{xy}}{S_x S_y}$ yields $r^2/1-r^2 = \frac{S_{xy}^2}{S_x^2 S_y^2 - S_{xy}^2}$

and hence

$$\frac{\hat{\rho}^2}{1-\hat{\rho}^2} = \frac{n\sigma_y^2}{S_y^2}\left\{\frac{r^2}{1-r^2}\right\}.$$

This implies that a statistic with some similarity to t is

$$Z = \left(\frac{n-2}{n}\right)\left(\frac{\hat{\rho}^2}{1-\hat{\rho}^2}\right) = \left[(n-2)\left(\frac{r^2}{1-r^2}\right)\right]/[S_y^2/\sigma_y^2]$$

Now letting

$$V = (n-2)\ \frac{r^2}{1-r^2}$$

and $U = S_y^2/\sigma_y^2$ yields $Z = V/U$.

Note first that under H_0, r^2 is independent of $y_1, y_2, \ldots, y_n$ [see the derivation of the conditional distribution of r^2 given y in Hogg and Craig (1970) or Johnson and Kotz (1970)]. Therefore $(n-2)\ \frac{r^2}{1-r^2}$ is independent of S_y^2/σ_y^2, that is, V is independent of U. Also note that $V \sim F(1,n-2)$ and $S_y^2/\sigma_y^2 \sim \chi^2(n-1)$. Therefore the p.d.f. of the random variable Z is [see Mood, Graybill and Boes (1974), p. 187]

$$f_z(z) = \int_{-\infty}^{\infty} |u| f_u(u) f_v(zu)\,du$$

$$= \int_0^{\infty} u\ \frac{1}{\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{1}{2}\right)^{\frac{n-1}{2}} u^{\frac{n-3}{2}} e^{-u/2}$$

$$\times\ \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)}\left(\frac{1}{n-2}\right)^{1/2} \frac{(zu)^{1/2}}{\left\{1+\frac{1}{n-2}\,zu\right\}^{\frac{n-1}{2}}}\,du$$

$$= \frac{2^{\frac{1-n}{2}}\,(n-2)^{-1/2} z^{-1/2}}{\Gamma\left(\frac{1}{2}\right)\ \Gamma\left(\frac{n}{2}-1\right)} \int_0^{\infty} \frac{u^{\frac{n-2}{2}} e^{-u/2}}{\left\{1+\left(\frac{1}{n-2}\right)zu\right\}^{\frac{n-1}{2}}}\,du\ .$$

The integral does not have a closed form expression. It was evaluated using 15-point Laguerre integration and verified using Whittaker functions.

To find the critical point we seek c such that for a one-sided $\alpha = 0.05$ test

$$P[Z \geq c] = 0.10$$

and then by symmetry we have

$$P[\sqrt{n-2}\ \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}} \geq \sqrt{nc}] = P[Z_1 \geq \sqrt{nc}] = 0.05.$$

Using a numerical search routine, c was found to be 0.47717 and hence for the statistic Z_1 the critical value is 2.1844 for $n = 10$ and $\alpha = .05$ (one sided test). A Monte Carlo study with one thousand samples of size $n = 10$ with $\rho = 0$ exhibited an empirical type I error rate of exactly 0.05.

A Monte Carlo power study to estimate the power of this test based on Z_1, along with the empirical powers of the other two competing tests based on t and t* is reported in Table 2. The table also presents the true powers of t (10,10) and t* (10,∞). There is no evidence that the test based on Z_1 is more powerful than the one based on t or t* for all alternative values of ρ. In fact, Z_1 has significantly less power for $0.1 \leq \rho \leq 0.5$. Figure 1 displays the power curves of the three tests. The fact that the test based on t* is the most powerful of the three near $\rho = 0$ is not discernible; yet even though we noted in Section III that t* was inferior to t for large ρ, it can be shown that t* is a locally most powerful invariant test [Ahmad and Giri (1979)]. It is also clear that this local optimality is rather inconsequential in light of the price

TABLE 2

Power of tests of H_o: $\rho = 0$ vs. H_1: $\rho > 0$ ($\alpha = 0.05$)

Monte Carlo Estimates Based on 2500 samples

	True Power	Empirical Power			True Power
ρ	$t(10,10)$	$t(10,10)$	$Z_1(10,\infty)$	$t^*(10,\infty)$	$t^*(10,\infty)$
.00	.050	.054	.055	.053	.050
.10	.085	.089	.079	.096	.0874
.20	.138	.147	.116	.153	.1418
.30	.215	.224	.188	.218	.2153
.40	.321	.312	.277	.307	.3087
.45	.390	.381	.325	.368	.3624
.50	.459	.441	.387	.415	.4202
.60	.623	.618	.584	.536	.5450
.75	.867	.867	.874	.739	.7385
.90	.993	.994	.999	.905	.9033

Maximum Standard Error of Table Entries: 0.010

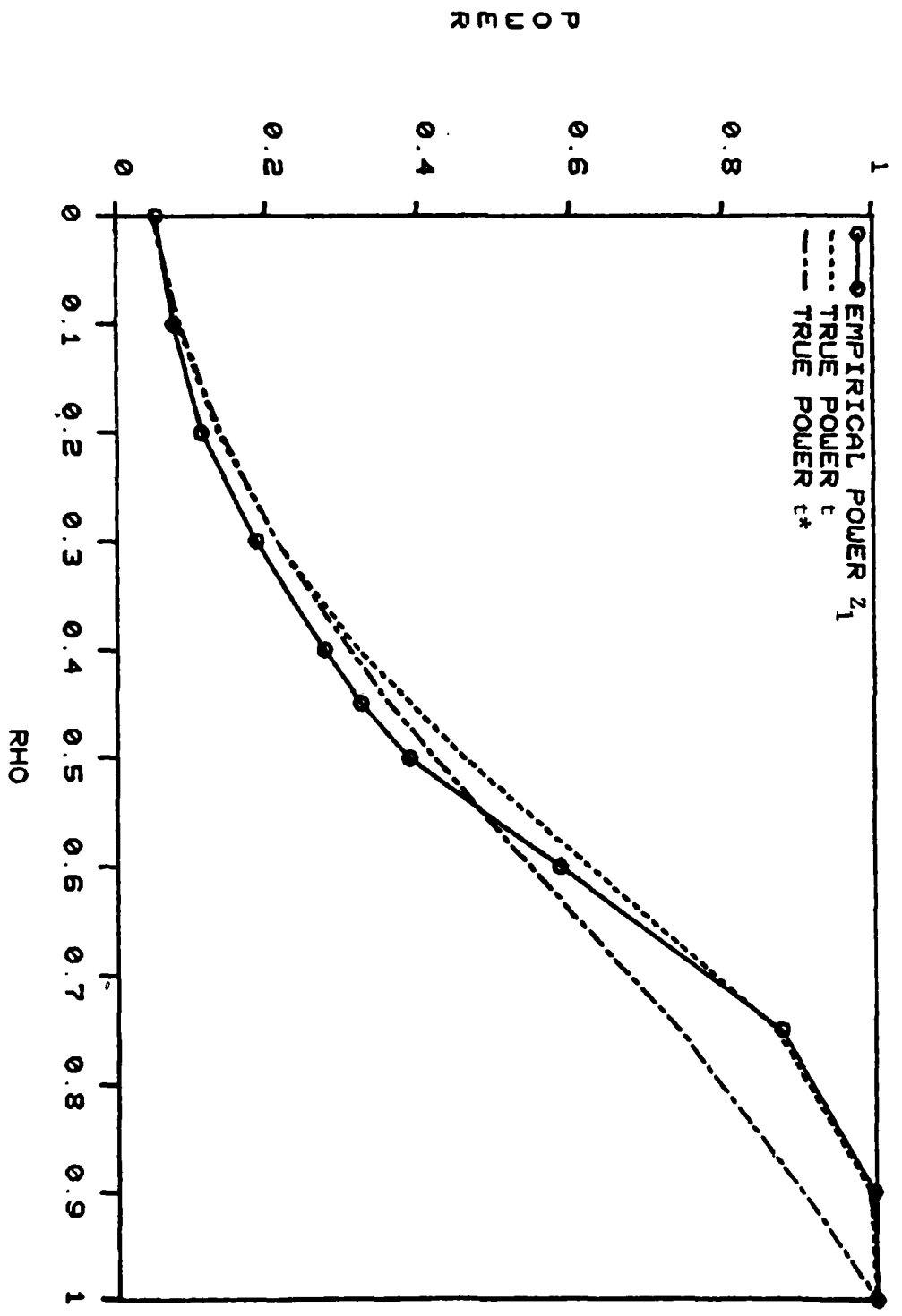


FIGURE 1

POWER CURVES FOR THREE TESTS

that is paid at the larger values of ρ. Finally, it seems apparent from Figure 1 that the usual t-test is preferable to either t* (LMP) or Z_1(MLE).

V. The Generalized Likelihood Ratio Test

For completeness it is interesting to obtain the generalized likelihood ratio statistic as a final competitor to the three tests of the previous section. Consider again the situation in which we have m additional observations on Y.

Following Anderson (1957), we let

$$\nu = \mu_x - \beta_{xy}\mu_y$$

$$\beta_{xy} = \rho\sigma_x/\sigma_y$$

$$\sigma^2_{x\cdot y} = \sigma^2_x(1-\rho^2)$$

and the likelihood function may be written

$$L = \prod_{i=1}^{n} \phi(x_i, y_i | \mu_x, \mu_y, \sigma^2_x, \sigma^2_y, \rho) \prod_{j=1}^{m} \phi(y_{n+j} | \mu_y, \sigma^2_y)$$

$$= \prod_{i=1}^{n} \phi(x_i | \nu + \beta_{xy} y_i;\ \sigma^2_{x\cdot y}) \prod_{j=1}^{n+m} \phi(y_i | \mu_y, \sigma^2_y) \ .$$

To maximize L the second product above is maximized whether H_0 is true or not at

$$\hat{\mu}_y = \frac{1}{n+m} \sum_{j=1}^{n+m} y_i$$

and

$$\hat{\sigma}^2_y = \frac{1}{n+m} \sum_{j=1}^{n+m} (y_j - \hat{\mu}_y)^2 \ .$$

Thus the likelihood ratio will depend only upon the two maxima of the products of the conditional densities of the x_i.

Under the null hypothesis $\rho = 0$, which implies that $\beta = 0$ and that the MLEs are

$$\hat{\nu} = \bar{x}_n$$

and

$$\hat{\sigma}^2_{x \cdot y} = S_x^2/n \quad .$$

The unconstrained maximization yields the usual estimates of regression parameters,

$$\hat{\beta}_{xy} = S_{xy}/S_{xy}^2 \ ,$$

$$\hat{\nu} = \bar{x}_n - \hat{\beta}_{xy}\bar{y}_n$$

and

$$\hat{\sigma}^2_{x \cdot y} = \frac{1}{n}(S_x^2 - \hat{\beta}_{xy}S_y^2)$$

$$= (S_x^2 S_y^2 - S_{xy}^2)/nS_y^2.$$

It follows algebraically that

$$\Lambda = \frac{L(H_0)}{L(\overline{H}_0)} = (1-r^2)^{n/2} \quad .$$

Notice that this does not depend upon the additional observations and offers still further support for the use of the familiar t statistic.

V. Summary

An exact test on the correlation coefficient which uses all the available data has been derived. In spite of the increase in degrees of freedom, the test cannot be recommended over the usual t test based only on the paired observations. The exact test is actually locally

most powerful invariant but this advantage quickly disappears for moderately large alternative values of ρ. The test based on the MLE of ρ also uses all of the data but is not better than the t test based only on the paired observations. Finally, the generalized likelihood ratio test is seen to be equivalent to the familiar t test; the unpaired observations are ignored. Thus in spite of the fact that t* 1) provides an exact test, 2) has greater degrees of freedom, 3) uses all the data and 4) is locally most powerful invariant -- all of which are desirable qualities -- it is inferior (practically speaking) to the familiar t test, which is based only on the paired observations.

REFERENCES

Ahmad, M. and Giri, N.C. (1979). "A test of bivariate independence with additional data", unpublished manuscript.

Anderson, T.W. (1957). "Maximum likelihood estimates for a multivariate normal distribution when some observations are missing," J. Amer. Statist. Assoc., 52, 200-203.

David, F.N. (1954). Tables of the Correlation Coefficient, Cambridge University Press, London.

Hogg, R.V. and Craig, A.T. (1970). Introduction to Mathematical Statistics, 3rd Edn., The MacMillan Co., Toronto.

Johnson, N. L. and Kotz, S. (1970). Continuous Univariate Distributions-2, Houghton Mifflin Co., Boston.

Mood, A., Graybill, F.A. and Boes, D.C. (1974). Introduction to the Theory of Statistics, 3rd Edn., McGraw-Hill, New York.

Wilks, S.S. (1932), "Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples," Annals of Math. Stat. 3, 163-195.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 153	2. GOVT ACCESSION NO.: AD-A109548
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): Testing the Correlation Coefficient with Incomplete Observations	5. TYPE OF REPORT & PERIOD COVERED: TECHNICAL REPORT
	6. PERFORMING ORG. REPORT NUMBER: 153
7. AUTHOR(s): James Beckett, N. J. Bosmia, William R. Schucany	8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0439
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Southern Methodist University, Dallas, Texas 75275	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Arlington, VA 22217	12. REPORT DATE: November 1981
	13. NUMBER OF PAGES: 17
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report):
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any purposes of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Correlation is often investigated (and tested for significance) in situations where some of the observations on one of the variables are missing. Throwing away these unpaired observations may seem to be a waste of information; a test based on all the data at hand would seemingly be better than a test based on only some of the data available. An exact test using all the data, which is similar in form and distribution to the usual t test based on the sample correlation coefficient, is derived and examined. However, this exact test proves to be a relatively inefficient way to incorporate the extra information. This counterintuitive result provides an interesting lesson concerning the relationship between power and degrees of freedom.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)